



## RESEARCH LETTER

10.1002/2016GL072012

## Key Points:

- Model weighting can constrain future projections
- Ensemble projections must also account for model interdependence
- Finding appropriate metrics to weight models remains challenging

## Correspondence to:

R. Knutti,  
reto.knutti@env.ethz.ch

## Citation:

Knutti, R., J. Sedláček, B. M. Sanderson, R. Lorenz, E. M. Fischer, and V. Eyring (2017), A climate model projection weighting scheme accounting for performance and interdependence, *Geophys. Res. Lett.*, *44*, 1909–1918, doi:10.1002/2016GL072012.

Received 18 NOV 2016

Accepted 6 FEB 2017

Accepted article online 8 FEB 2017

Published online 18 FEB 2017

## A climate model projection weighting scheme accounting for performance and interdependence

Reto Knutti<sup>1,2</sup> , Jan Sedláček<sup>1</sup> , Benjamin M. Sanderson<sup>2</sup> , Ruth Lorenz<sup>1</sup> , Erich M. Fischer<sup>1</sup> , and Veronika Eyring<sup>3</sup>
<sup>1</sup>Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland, <sup>2</sup>National Center for Atmospheric Research, Boulder, Colorado, USA, <sup>3</sup>Deutsches Zentrum für Luft- und Raumfahrt (DLR), Institut für Physik der Atmosphäre, Oberpfaffenhofen, Germany

**Abstract** Uncertainties of climate projections are routinely assessed by considering simulations from different models. Observations are used to evaluate models, yet there is a debate about whether and how to explicitly weight model projections by agreement with observations. Here we present a straightforward weighting scheme that accounts both for the large differences in model performance and for model interdependencies, and we test reliability in a perfect model setup. We provide weighted multimodel projections of Arctic sea ice and temperature as a case study to demonstrate that, for some questions at least, it is meaningless to treat all models equally. The constrained ensemble shows reduced spread and a more rapid sea ice decline than the unweighted ensemble. We argue that the growing number of models with different characteristics and considerable interdependence finally justifies abandoning strict model democracy, and we provide guidance on when and how this can be achieved robustly.

## 1. Motivation

Our ability to predict climate change is riddled with uncertainties: different scenarios of societal and technological development, imperfect models, and natural variability are the main reasons that future projections cannot be certain and deterministic [Knutti, 2008; Hawkins and Sutton, 2009; Knutti and Sedláček, 2012]. Scenario uncertainty is accounted for by making projections (rather than predictions) conditional on an emission scenario. Natural variability is an inherent component of any evolution of climate and is nearly irreducible for lead times greater than 5–10 years [Deser et al., 2012]. For projections, we are left mostly with uncertainty arising from our incomplete understanding of nature and its representation in models. For large-scale long-term projections, this is often the dominant source of uncertainty. A common approach to quantify it is to consider multiple models from different institutions running common experiments, such as in the Coupled Model Intercomparison Projects (CMIP) [Eyring et al., 2016], but each group making different but plausible choices regarding which processes are included and how they are represented [Parker, 2006; Tebaldi and Knutti, 2007; Knutti, 2008; Sanderson and Knutti, 2012]. The prevailing approach for dealing with multimodel ensemble results is model democracy [Knutti, 2010]. Often based on a lack of convincing alternatives, this “one model one vote” essentially assumes that all models are (a) reasonably independent, (b) equally plausible, and (c) distributed around reality and (d) that the range of their projections is representative of what we believe is the uncertainty in the projected quantity.

Strictly speaking, none of the four conditions is fulfilled by model democracy. On (a), many models duplicate ideas, or even use large parts of the code of others, so at best are providing little additional information and at worst are biasing the result [Annan and Hargreaves, 2011; Masson and Knutti, 2011a; Pennell and Reichler, 2011; Knutti et al., 2013]. On (b), some models are worse than others in how well they represent the observed mean climate and trends [Gleckler et al., 2008; Reichler and Kim, 2008; Knutti et al., 2013]. On (c) models have common structural limitations so the ensemble as a whole may be biased. And on (d), our suite of models may be either too broad if we have demonstrably unrealistic models included (e.g., one of Venus) or too narrow if all models are missing the same processes, or make similar approximations. We do not know which of the two is correct, and this may differ between variables, regions, and time scales as model performance is strongly scale and variable dependent [Masson and Knutti, 2011b]. We expect that equal weights for all models are a suboptimal way of using information, yet this remains common, for the lack of better and easy alternatives or consensus on what those are. We provide evidence below that for some applications (and to the degree that the perfect model test is informative) there are better alternatives.

The arguments against weighting models are the following: The first is that we do not know how to weight. While it is easy to define a “model performance metric” (e.g., the difference between simulated and observed rainfall), we do not know or agree on how to transfer this into a “model quality metric” that is indicative of skill for a projection [Knutti *et al.*, 2010b], and eventually into weights. The situation is more challenging than, for example, for weather forecasts where we can quantify skill by repeated verification. Confidence in climate projections must derive from developers’ understanding of how the climate system works, whether the relevant processes are captured in the model or how that manifests itself in the simulation. That degree of confidence is partly subjective. The second argument often made against weighting is robustness. Betting on a few or only one model is risky and may lead to a biased or overconfident result [Weigel *et al.*, 2010], but that is only a major issue if the number of models is small or if most of the weight is put on a few models and when the weight is unrelated to the projected variable. Finally, there is a political element in that eliminating models from an ensemble is judged by some to be inappropriate in international assessments, or there is no consensus on how to do it.

A way out, at least partially, is to reframe the competition of “which model is the best” to “which models are adequate for predicting  $X$ ” [Parker, 2009]. The former is an ill-posed question; there is no best model without defining what “better” means. But it is easier to define criteria for a model to be better or worse for a particular purpose. If we define weights for predicting  $X$  or  $Y$ , then the situation becomes easier both scientifically and politically. One model may get more weight at predicting  $X$ , and another one at predicting  $Y$ , which is natural as different institutions focus on different questions. A model would only be downweighted for all purposes if, for example, it strongly violates conservation of water or energy. Defining weights for predicting  $X$  is also easier because we have an idea of which processes are important for  $X$ . In some cases, the ensemble can be used to find relationships between observable aspects of mean climate, variability, or trends and the prediction of  $X$ , or some feedbacks affecting it. This idea of “emergent constraints” is now explored widely.

Methods beyond democracy include statistical methods that introduce some statistical metamodel, e.g., based on regression between observables and predictions using emergent constraints [Hall and Qu, 2006; Boé *et al.*, 2009a; Huber *et al.*, 2011; Mahlstein and Knutti, 2012; Cox *et al.*, 2013; Fasullo *et al.*, 2015], interpolation in a model space [Sanderson *et al.*, 2015b], Bayesian methods [e.g., Tebaldi *et al.*, 2004], or combinations thereof. The results are often relatively independent of the underlying sample, and in some cases extrapolation beyond the model range and probabilistic estimates are possible. Some studies have assumed a problematic “truth plus error” paradigm, i.e., that models are independent and distributed around reality (for an in-depth discussion see Knutti *et al.* [2010b], Annan and Hargreaves [2011], and Sanderson and Knutti [2012]), which leads to overly narrow results for a large number of models [Knutti *et al.*, 2010a]. The emergent constraints may be overestimated if all models have structurally similar problems or the ensemble is too small [Caldwell *et al.*, 2014]. Conservation of quantities and consistency across time, space, and across variables is often lost for multivariate results. Finally, the technical and methodological hurdles have restricted the use of some methods.

An alternative to regressions across models is picking subsets or weighting individual model simulations. This can provide multivariate data sets that are consistent across variables, time, and space (to the degree that the model reflects reality) but allows no straightforward extrapolation beyond the model range. Most methods in this category consider performance but ignore model dependence, and those which do are technically challenging [Abramowitz and Bishop, 2015; Sanderson *et al.*, 2015a].

We argue that weighting schemes that consider model performance and interdependence are required for at least five reasons. First, for some regions and variables the model spread in the present day climatology is massive, and thereby biases in some models are so large that model democracy is difficult to justify. For example, the output of a model in which Arctic sea ice has already disappeared today cannot be used directly to predict future Arctic sea ice (although its sensitivity of sea ice to temperature could be used potentially, which would be another way of weighting by an emergent constraint) [Boé *et al.*, 2009b; Mahlstein and Knutti, 2012; Notz and Stroeve, 2016]. Second, in cases where processes are sensitive to the base state (e.g., temperature determining the sea ice edge or temperature variability change depending nonlinearly on soil moisture), working with projected anomalies relative to today is problematic, and scaling methods and bias correction may be unfeasible. Third, in some cases there are emergent constraints that are clearly relevant for model evaluation and that can improve projections (as in the sea ice example discussed here),

so it would be strange not to use them. Fourth, large initial condition ensembles (or other types of simulations) are hard to combine with single runs from other models. And fifth, model dependence gets increasingly relevant with increasing replication of code across institutions, models being run at different resolutions, used as climate versus Earth system models, etc. [Knutti *et al.*, 2013]. This approach of sharing ideas and code is natural to avoid duplication of efforts, but the dependence has to be considered in the interpretation.

## 2. Method

The basic idea of the method presented here is simple: models that agree poorly with observations for a selected set of diagnostics get less weight and models that largely duplicate existing models also get less weight. The proposed scheme therefore extends previous approaches on weighting multimodel projections of for example, sea ice [Massonnet *et al.*, 2012] or stratospheric ozone [Vaugh and Eyring, 2008] by additionally considering model interdependence. Note that this method is limited to weighting maps or time series and does not straightforwardly apply to other methods for making projections.

Any weighting scheme inevitably requires making important and subjective choices on the distance metric, its conversion into weight, diagnostics and observations (including their uncertainties) to be used, and the relative importance of the different diagnostics. Some of these choices are not straightforward at all and need to be defined and assessed specifically for each application.

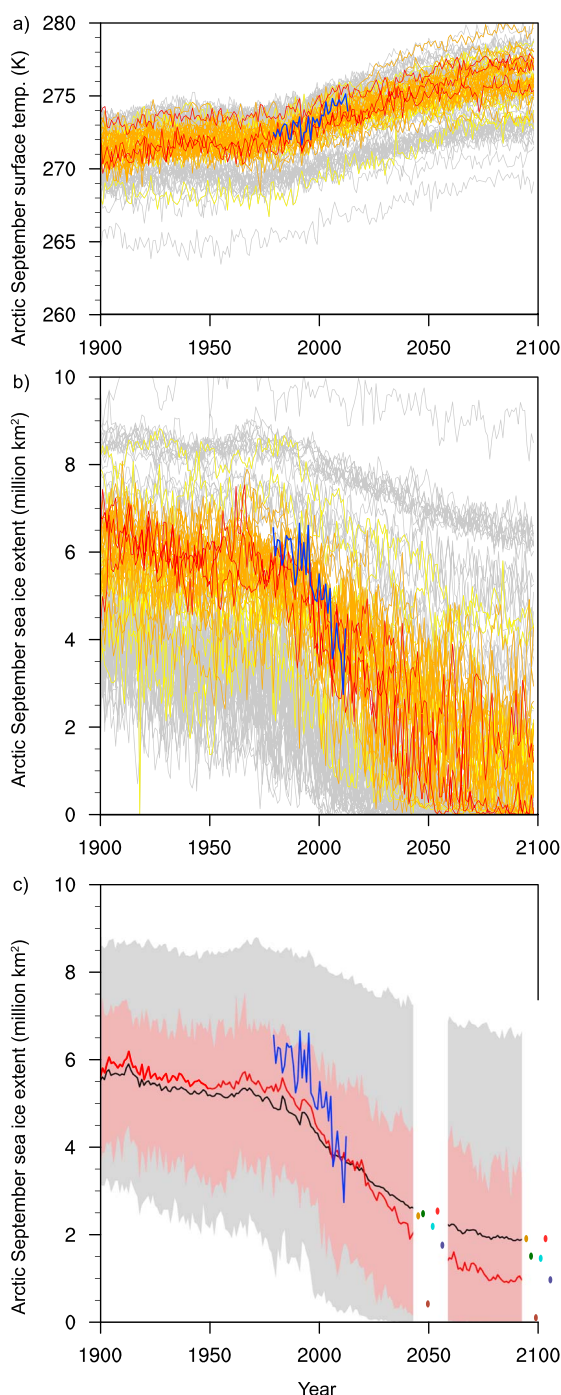
The weighting first requires defining a distance metric  $D_i$  of model  $i$  to observations, and  $S_{ij}$ , the distance metric between model  $i$  and model  $j$ , and a relationship to convert those into a weight. Both  $D_i$  and  $S_{ij}$  are evaluated in our example as root-mean-square differences, but other metrics are also possible. For  $M$  models in the ensemble, the single model weight  $w_i$  for model  $i$  is defined as follows:

$$w_i = e^{-\frac{D_i^2}{\sigma_D^2}} / \left( 1 + \sum_{j \neq i}^M e^{-\frac{S_{ij}^2}{\sigma_S^2}} \right). \quad (1)$$

The constants  $\sigma_D$  and  $\sigma_S$  determine how strongly the model performance and similarity are weighted and are discussed below. The weights are normalized so that their sum equals one. The interpretation of the weighting is straightforward, based on and further explained by Sanderson *et al.* [2015a, equations (10)–(16)]. The scheme here is identical except that no empirical orthogonal functions are used, and weights are used directly rather than for interpolation in a model space or producing ensemble subsets. The numerator represents model skill by using a Gaussian weighting where the weight decreases exponentially the further away a model is from observations. The denominator is the “effective repetition of a model” [Sanderson *et al.*, 2015a] and is intended to account for model interdependency. If a model has no close neighbors, then all  $S_{ij}$  ( $i \neq j$ ) are large, the denominator is approximately one and has no effect. If two models  $i$  and  $j$  are identical, then  $S_{ij} = 0$ , the denominator equals two, so each model gets half the weight.

This scheme fulfills the two main requirements: a model that bears no similarity to Earth ( $D_i$  being very large) gets no weight and adding a model that is identical to one already there does not change the result. This scheme also deals naturally with multiple initial condition ensembles; all members can be used, even if not all models provide the same number of ensembles, and the method effectively treats initial condition members as duplicate models and downweights them accordingly. The older CMIP3 models can be included as well [Rauser *et al.*, 2015] if the scenario is sufficiently similar and if they are worse than newer models they simply get less weight. Ensemble spread is calculated as a weighted standard deviation.

Second, the metrics  $D_i$  and  $S_{ij}$  require a choice of the diagnostics and variables that are relevant for the projection of a certain variable. When several diagnostics are used, they are normalized here by the median distance across all models/members before adding them to the total distance, so that each diagnostic gets about the same weight. But many metrics and ways to combine variables are possible, and choices can be made either based on expert judgment about relevant processes, on emergent relationship across models, or multiple diagnostic ensemble regression methods [Vaugh and Eyring, 2008; Karpechko *et al.*, 2013; Sanderson *et al.*, 2015b; Wenzel *et al.*, 2016a]. Observational uncertainty, significance of differences, and sensitivities to the choice of data sets need to be tested. In the example here, the temperature and sea ice biases in models, however, are far larger than observational uncertainties.



**Figure 1.** (a) Arctic (60–90°N) September surface air temperature and (b) Arctic September sea ice extent in all CMIP3/5 simulations. Yellow, orange, and red indicate those that get >0.5%, >1%, and >5% weight, respectively, from weighting with equation (1). Observations (National Centers for Environmental Prediction, NCEP) are shown in blue. (c) Mean and 5–95% range for no weighting (black line and grey band) and weighting (red line and band). Colored dots near 2050 and 2100 show 2046–2055 and 2090–2099 average sea ice extent using (from left to right) the following metrics: (1) none (unweighted), (2) climatological mean (1980–2013) September sea ice extent, (3) September sea ice extent trend 1980–2013, (4) climatology of monthly surface temperature (1980–2013), (5) interannual variability of monthly surface temperature, and (6) all 2–5.

The choice of  $\sigma_D$  and  $\sigma_S$  determines how strongly the model performance and similarity are weighted. A large  $\sigma_D$  effectively converges to model democracy, whereas a small  $\sigma_D$  puts the weight on only a few models;  $\sigma_S$  determines a typical distance by which a model would be considered similar to another one. The choice of those values is discussed along with the results. A more formal way of treating dependence would be desirable, but the conceptual ideas being discussed are not applicable in an obvious way [Annan and Hargreaves, 2016]. Dependence and performance are treated independently here, and one concern may be that two independent models converging to reality become more similar in our definition of  $S$  and thus might be penalized unjustly. A similarity metric based on correlation [Watterson, 1996] would eliminate this but would lose information on the absolute distance. For the metrics considered here, the typical distance to observation is large compared to the distance between duplicate models, and the results are rather insensitive to how strongly model dependence is weighted. That should alleviate concerns, but these questions will require further conceptual work and testing in various applications.

### 3. Application to Arctic Sea Ice and Temperature Projections

To demonstrate the application and skill of the proposed method, we consider projections of Arctic September temperature and sea ice. Figures 1a and 1b show time series of absolute Arctic mean temperature and total sea ice extent for each CMIP5 simulation (historical and Representative Concentration Pathway (RCP4.5), all initial condition members), respectively. CMIP3 simulations with the SRES B1 scenario are also included. While not identical, B1 and RCP4.5 are similar enough to be analyzed jointly [Knutti and Sedláček, 2012].

Weights are assigned based on the above equation (1). The default performance and similarity metric combines distances in the following four diagnostics: the climatological mean hemispheric mean September Arctic sea ice extent (1980–2013), its trend over the same period, gridded climatological mean surface air temperature for each month, and climatological mean gridded interannual variability of monthly surface air temperature. Distances are aggregated as root-mean-square differences. The arguments for using those diagnostics for weighting Arctic temperature and sea ice projections are as follows: some models have almost no sea ice today, and others have more sea ice in 2100 than observed today, and therefore, they are not suitable for a projection of future sea ice. Absolute temperature biases are large in some simulations. There are clear relationships across models between present-day and future polar amplification [Bracegirdle and Stephenson, 2012, 2013], as well as between past and future sea ice trends and temperature [Boé et al., 2009b; Mahlstein and Knutti, 2012; Massonnet et al., 2012; Overland and Wang, 2013; Notz and Stroeve, 2016]. A stronger ice albedo feedback also explains more rapid sea ice loss in both the past and future, which is a plausible physical explanation for those relationships. But since sea ice extent depends nonlinearly on absolute temperature (and vice versa) once all sea ice has disappeared, most simple scaling or bias correction methods based on anomalies fail, while weighting produces a projection that is consistent with and calibrated by past observations. In addition to matching the mean and trend in sea ice, we also evaluate the seasonal and spatial patterns of temperature and its variability to ensure the sea ice match is not by chance but a result of a decent representation of the overall Arctic climate.

Those simulations with nonnegligible weights are shown in colors in Figures 1a and 1b, with yellow to red indicating less to more weight. The red shaded band in Figure 1c shows the 5–95% confidence interval based on the weighted ensemble. The weighted projection points to near ice-free September conditions by 2100 for RCP4.5, with a very likely range of zero to about  $4 \times 10^6 \text{ km}^2$ . That result is more consistent with, and better constrained by, the current extent and past trends of sea ice than the grey range of the full ensemble which is inconclusive.

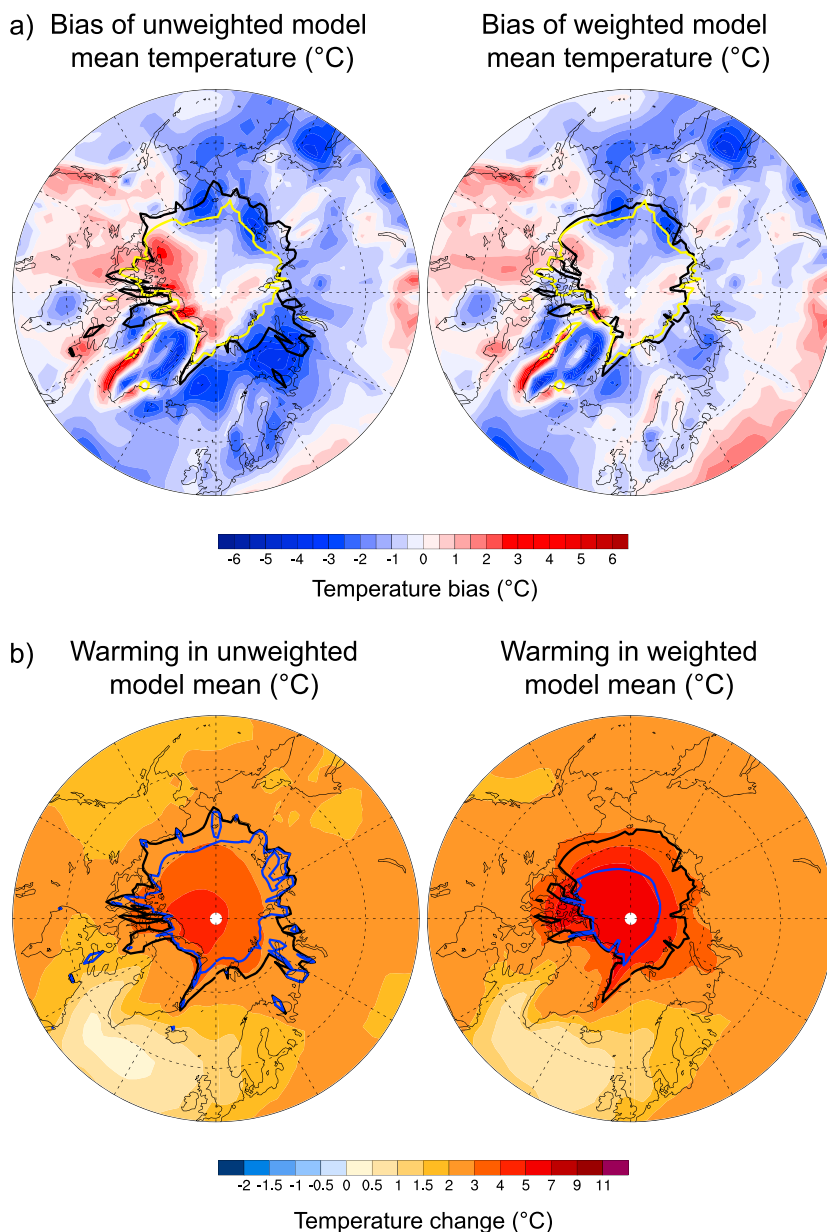
The results are robust to picking different individual metrics from the four diagnostics, in that the calibrated projections always show either no substantial difference or a tendency toward faster sea ice decline than the unweighted case (see colored dots for different metrics in Figure 1c), and a strong reduction in spread (not shown). The calibrated projections demonstrate an improved agreement in the present-day model mean surface temperature (Figure 2a). They also project more warming over the central Arctic (up to  $2^\circ\text{C}$ ) compared to the unweighted case (Figure 2b), and a faster sea ice decline, consistent with other calibrated estimates [Boé et al., 2009b; Mahlstein and Knutti, 2012; Massonnet et al., 2012; Overland and Wang, 2013].

Another approach to evaluate the method is a perfect model setup (or cross validation or pseudoreality) [Karpechko et al., 2013; Wenzel et al., 2016a], where each model is sequentially treated as “truth” and the others are weighted to predict its future response. Figure 3a shows an example of the high correlation between predicted and true sea ice extent. Figure 3b shows that this correlation is high for different choices of  $\sigma_D$ . The dependence on the value of  $\sigma_S$  is rather weak relative to  $\sigma_D$  (not shown). Figure 3b would favor a low value for  $\sigma_D$  to maximize correlation, and indeed, that would further improve the agreement with observed sea ice trends and result in a more rapid decline of Arctic sea ice. However, correlation only tells us about the projected best estimate, is insensitive to constant biases, and does not consider the projection uncertainty.

As an alternative metric, Figure 3c shows the fraction of cases when the actual outcome of the perfect model test is in the 5–95% range predicted by weighting the other models. That fraction should be around 90%, which points to a minimum value for  $\sigma_D$  of about 0.37 used here. Any value lower than that leads to overconfident results within the ensemble and therefore is likely to produce overly narrow projections, even if the agreement in the observational period is good. Note that part of the high correlation may result from multiple initial condition members or duplicate models providing an unrealistically good match with other models. Using only one model per institution suggests a value for  $\sigma_D$  of 0.57, which would result in future uncertainty ranges being about 20% wider. But on the other hand, that may be overly conservative, as it forces the method to weigh less restrictive to capture models with extreme behavior.

There is no objective way to best determine these parameters, but conclusions are robust for different choices. The comparison of the weighted average and observed evolution (Figure 1c, red and blue lines)

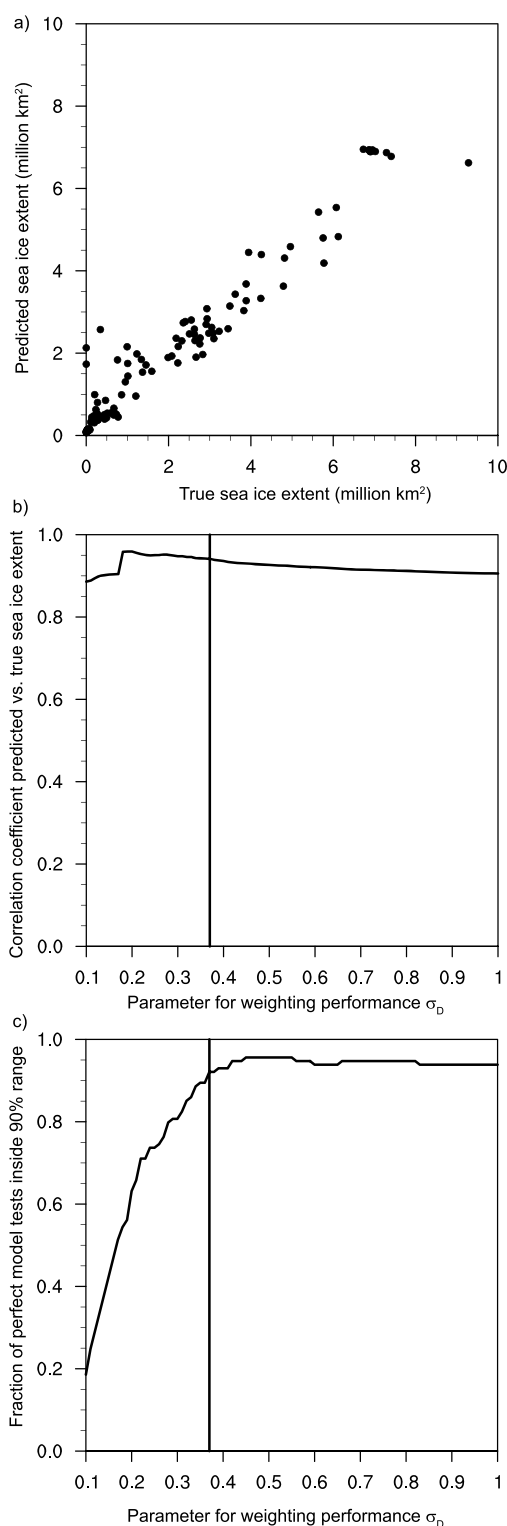




**Figure 2.** (a) Unweighted and weighted model mean September surface temperature bias relative to NCEP 2000–2009. The black line indicates the modeled September sea ice extent; yellow indicates the observed sea ice extent. (b) Unweighted and weighted model mean September surface warming 2085–2094. The black and blue lines indicate the present-day and future September sea ice extent, respectively.

thus suggests that part of the recent rapid Arctic warming and observed decrease in sea ice may be due to natural variability [Kay *et al.*, 2011; Swart *et al.*, 2015; Screen and Francis, 2016]. Simply extrapolating past trends or calibrating models on a too narrow metric, e.g., only on the observed sea ice trend (Figure 1c, brown points, case 3), may overestimate the forced trend in this case (or underestimate it if natural variability had damped the forced response) and be unreliable.

Such perfect model setups can be used to test the skill of the method, flag overfitting to natural variability or certain data sets, and quantities such as those in Figure 3 can help guiding the choice of parameters and metrics (in particular,  $\sigma_D$  and  $\sigma_S$  in this case). Such tests will also flag cases where the observations do not provide a useful constraint on the projection (which is often the case), and the method would then give similar weights to most models. Remaining questions are whether the choice of parameters or metrics should be



**Figure 3.** (a) Correlation between predicted and true September sea ice extent for each model in a perfect model setup using the standard values of  $\sigma_D = 0.37$  and  $\sigma_S = 0.5$ . (b) Dependence of the correlation shown in Figure 3a on  $\sigma_D$  and (c) fraction of cases when the actual outcome of the perfect model test is in the 5–95% range predicted by weighting all other models. Vertical line indicates the minimum value for  $\sigma_D$  for the fraction to exceed the required 90%.

influenced by the ability of the weighting to reproduce the results of an instantly dismissible model. Inference from such perfect model tests is limited if the emergent constraint is artificially high due to common structural model biases, large amounts of replication, or too few models. While open questions remain, such perfect model tests are a necessary test to pass, and we should have more confidence in the results with them than without them.

## 4. Conclusions

We presented a weighting scheme for multimodel climate model projections that considers both model performance and interdependence and illustrated its application to Arctic climate. We highlight several open questions but argue that in cases where obvious model performance criteria exist, there are schemes to weight model projections that are very likely better than treating all models equally, for both predicting a model mean and for estimating an uncertainty. A weighted climate model ensemble can be used as input for an impact model, the only difference is a single weight attached to each simulation, although the weights will be application specific. The scheme proposed naturally deals with multiple initial condition ensemble members and downweights obvious duplicate models. Giving more weight to better or newer models does not necessarily reduce the uncertainty in a projection [Knutti and Sedláček, 2012], but it increases our confidence when results are based on models that simulate relevant aspects of current climate more realistically. Good agreement with observations is not a proof that the model is correct [Baumberger et al., 2017], but bad agreement is a clear indication for trouble. A much stronger argument for such a method to work would be a true out of sample evaluation, e.g., by using data only up to say 1990 to constrain projections for 2020 [Allen et al., 2013]. In practice this is rarely possible because (a) the data up to 1990 are short, of too poor quality or has too much variability, (b) the forecast lead time is too short for a forced signal to emerge from variability, (c) data since 1990 are known and used

already in model development and evaluation, and (d) one realization is insufficient to assess skill in the presence of variability.

There are some stumbling blocks in methods such as the one outlined here. The critical issue is the choice of the diagnostics and their respective weights. They must depend on and be relevant for the quantity  $X$  being projected and can include agreement of trends, patterns, spectra, and information about processes related to feedbacks, indeed anything that we judge being important for  $X$ . While differences to observations are the most obvious choice, other criteria may also be justified. Substantial problems in conserving energy or mass, or drift in a control run, could be reasons to downweight or exclude models. Also, the degree of sophistication with which relevant processes are implemented (e.g., simple bucket-type land hydrology versus sophisticated land surface, hydrology, soil and runoff schemes, or interactive aerosol microphysics and stratospheric ozone chemistry versus prescribed distributions) can be considered. Methods have been proposed that help deciding which diagnostics matter most to weight a specific projection [Karpechko *et al.*, 2013; Wenzel *et al.*, 2016a] or that can directly constrain aspects of simulated future Earth system feedback with observations [e.g., Hall and Qu, 2006; Cox *et al.*, 2013; Sherwood *et al.*, 2014; Wenzel *et al.*, 2014, 2016b]. Other research directions are more sophisticated weighting frameworks, the interpretation and combination of perturbed physics and multi model ensembles [Knutti *et al.*, 2010b; Annan and Hargreaves, 2011, 2016; Sanderson and Knutti, 2012], and methods to combine, for example, regional models [Zubler *et al.*, 2015].

The choice of  $\sigma_D$  and  $\sigma_S$  is important. Pragmatic criteria for  $\sigma_D$  are that the weights should be distributed on more than just a few models. Results that are sensitive to the choice of data sets and its uncertainty, time period, the set of models or metrics considered indicate too aggressive weighting (too small  $\sigma_D$ ).

In any application of model weighting, we argue that authors must (a) show unweighted along with weighted results, (b) test the robustness of the results toward different diagnostics and metrics to maximize transparency and comparability across studies, (c) explicitly discuss the choice of diagnostics, including the physical reasoning that those quantities matter and possibly a formal framework to define the individual weights, (d) assess the uncertainties in observations, (e) test the sensitivity toward different data sets, time periods, seasonal versus annual mean values, grid point versus spatially aggregated data, etc., and (f) explore whether the choice of metric may lead to overconfident results, for example, by using the perfect model approach. If the results are not robust toward any of those choices, this indicates that weighting may be too aggressive (overfitting). Even though the method presented here is conceptually simple, its application requires understanding of the relevant processes and a careful choice of the diagnostics and parameters. Often both aggressive weighting and democracy are not optimal, and a sweet spot is somewhere in between. There are still many cases where the relevant diagnostics for a particular projection are largely unclear and where weighting has to be applied cautiously if at all.

While model performance has been a topic for a long time, the model replication/dependence issue has not received much attention. Replication will likely get worse in the future: since only few groups can afford to develop all model components, they are using components of other models, or develop them jointly. The number of submitted models will likely increase for the next model intercomparison, but the effective number of independent models might not. Future assessments will have to consider that.

Scientists are often concerned about weighting with inappropriate diagnostics. However, if the diagnostics used are indeed unrelated to the projection, then all models are essentially assigned random weights, so the results should not change much, unless the number of models is small [Weigel *et al.*, 2010]. There may be many cases where we do not know which diagnostics matter, and then reverting back to democracy may be the safest option. But we should be just as concerned about not weighting when we know that poor models are biasing our results or when models are replicated many times in the ensemble. The proposed weighting scheme here is just one option, but the main point is that such a scheme can be further explored for different applications and may improve projections. In the 2007 World Meteorological Organization/United Nations Environment Programme ozone assessment and the most recent Intergovernmental Panel on Climate Change assessment reports, total column ozone, sea ice, and near-term temperature trends were among the first projections that explicitly used observational constraints. Future research will help to extend that list to other cases where we can go beyond model democracy or arbitrary weighting. This will provide projections that are more consistent with observed present-day climate and past trends and therefore will be likely more reliable.



## Acknowledgments

We thank Nadja Herger for insightful comments and discussions. This project was supported by the European Union's Horizon 2020 research and innovation program under grant agreement 641816 (CRESCENDO) and by the Regional and Global Climate Modeling Program (RGCM) of the U.S. Department of Energy, Office of Science (BER), Cooperative Agreement DE-FC02-97ER62402. We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups for producing and making available their model output. For CMIP the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. CMIP data can be obtained from <http://cmip-pcmdi.llnl.gov/cmip5/>. The National Center for Atmospheric Research is sponsored by the National Science Foundation.

## References

- Abramowitz, G., and C. H. Bishop (2015), Climate model dependence and the ensemble dependence transformation of CMIP projections, *J. Clim.*, **28**, 2332–2348, doi:10.1175/JCLI-D-14-00364.1.
- Allen, M. R., J. F. B. Mitchell, and P. A. Stott (2013), Test of a decadal climate forecast, *Nat. Geosci.*, **6**(4), 243–244, doi:10.1038/ngeo1788.
- Annan, J., and J. Hargreaves (2016), On the meaning of independence in climate science, *Earth Syst. Dyn. Discuss.*, 1–17, doi:10.5194/esd-2016-34.
- Annan, J. D., and J. C. Hargreaves (2011), Understanding the CMIP3 multimodel ensemble, *J. Clim.*, **24**(16), 4529–4538, doi:10.1175/2011JCLI3873.1.
- Baumberger, C., R. Knutti, and G. Hirsch Hadorn (2017), Building confidence in climate model projections: An analysis of inferences from fit, *Wiley Interdiscip. Rev. Clim. Change*, e454, doi:10.1002/wcc.454.
- Boé, J., A. Hall, and X. Qu (2009a), Deep ocean heat uptake as a major source of spread in transient climate change simulations, *Geophys. Res. Lett.*, **36**, L22701, doi:10.1029/2009GL040845.
- Boé, J., A. Hall, and X. Qu (2009b), September sea-ice cover in the Arctic Ocean projected to vanish by 2100, *Nat. Geosci.*, **2**(5), 341–343, doi:10.1038/ngeo467.
- Bracegirdle, T. J., and D. B. Stephenson (2012), Higher precision estimates of regional polar warming by ensemble regression of climate model projections, *Clim. Dyn.*, **39**(12), 2805–2821, doi:10.1007/s00382-012-1330-3.
- Bracegirdle, T. J., and D. B. Stephenson (2013), On the robustness of emergent constraints used in multimodel climate change projections of Arctic warming, *J. Clim.*, **26**(2), 669–678, doi:10.1175/JCLI-D-12-00537.1.
- Caldwell, P. M., C. S. Bretherton, M. D. Zelinka, S. A. Klein, B. D. Santer, and B. M. Sanderson (2014), Statistical significance of climate sensitivity predictors obtained by data mining, *Geophys. Res. Lett.*, **41**, 1803–1808, doi:10.1002/2014GL059205.
- Cox, P. M., D. Pearson, B. B. Booth, P. Friedlingstein, C. Huntingford, C. D. Jones, and C. M. Luke (2013), Sensitivity of tropical carbon to climate change constrained by carbon dioxide variability, *Nature*, **494**(7437), 341–344, doi:10.1038/nature11882.
- Deser, C., R. Knutti, S. Solomon, and A. S. Phillips (2012), Communication of the role of natural variability in future North American climate, *Nat. Clim. Change*, **2**(11), 775–779, doi:10.1038/nclimate1562.
- Eyring, V., S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor (2016), Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, *Geosci. Model Dev.*, **9**(5), 1937–1958, doi:10.5194/gmd-9-1937-2016.
- Fasullo, J. T., B. M. Sanderson, and K. E. Trenberth (2015), Recent progress in constraining climate sensitivity with model ensembles, *Curr. Clim. Chang. Rep.*, **1**(4), 268–275, doi:10.1007/s40641-015-0021-7.
- Gleckler, P. J., K. E. Taylor, and C. Douriaux (2008), Performance metrics for climate models, *J. Geophys. Res.*, **113**, D06104, doi:10.1029/2007JD008972.
- Hall, A., and X. Qu (2006), Using the current seasonal cycle to constrain snow albedo feedback in future climate change, *Geophys. Res. Lett.*, **33**, L03502, doi:10.1029/2005GL025127.
- Hawkins, E., and R. Sutton (2009), The potential to narrow uncertainty in regional climate predictions, *Bull. Am. Meteorol. Soc.*, **90**(8), 1095–1107, doi:10.1175/2009BAMS2607.1.
- Huber, M., I. Mahlstein, M. Wild, J. Fasullo, and R. Knutti (2011), Constraints on climate sensitivity from radiation patterns in climate models, *J. Clim.*, **24**(4), 1034–1052, doi:10.1175/2010JCLI3403.1.
- Karpechko, A. Y., D. Maraun, and V. Eyring (2013), Improving Antarctic total ozone projections by a process-oriented multiple diagnostic ensemble regression, *J. Atmos. Sci.*, **70**(12), 3959–3976, doi:10.1175/JAS-D-13-071.1.
- Kay, J. E., M. M. Holland, and A. Jahn (2011), Inter-annual to multi-decadal Arctic sea ice extent trends in a warming world, *Geophys. Res. Lett.*, **38**, L15708, doi:10.1029/2011GL048008.
- Knutti, R. (2008), Should we believe model predictions of future climate change?, *Philos. Trans. R. Soc. London, Ser. A*, **366**(1885), 4647–4664, doi:10.1098/rsta.2008.0169.
- Knutti, R. (2010), The end of model democracy?, *Clim. Change*, **102**(3–4), 395–404, doi:10.1007/s10584-010-9800-2.
- Knutti, R., and J. Sedláček (2012), Robustness and uncertainties in the new CMIP5 climate model projections, *Nat. Clim. Change*, **3**(4), 369–373, doi:10.1038/nclimate1716.
- Knutti, R., R. Furrer, C. Tebaldi, J. Cermak, and G. A. Meehl (2010a), Challenges in combining projections from multiple climate models, *J. Clim.*, **23**(10), 2739–2758, doi:10.1175/2009JCLI3361.1.
- Knutti, R., G. Abramowitz, M. Collins, V. Eyring, P. J. Gleckler, B. Hewitson, and L. Mearns (2010b), Good practice guidance paper on assessing and combining multi model climate projections, edited by T. F. Stocker et al., Univ. of Bern, Bern, Switzerland.
- Knutti, R., D. Masson, and A. Gettelman (2013), Climate model genealogy: Generation CMIP5 and how we got there, *Geophys. Res. Lett.*, **40**, 1194–1199, doi:10.1002/grl.50256.
- Mahlstein, I., and R. Knutti (2012), September Arctic sea ice predicted to disappear near 2°C global warming above present, *J. Geophys. Res.*, **117**, D06104, doi:10.1029/2011JD016709.
- Masson, D., and R. Knutti (2011a), Climate model genealogy, *Geophys. Res. Lett.*, **38**, L08703, doi:10.1029/2011GL046864.
- Masson, D., and R. Knutti (2011b), Spatial-scale dependence of climate model performance in the CMIP3 ensemble, *J. Clim.*, **24**(11), 2680–2692, doi:10.1175/2011JCLI3513.1.
- Massonnet, F., T. Fichefet, H. Goosse, C. M. Bitz, G. Philippon-Berthier, M. M. Holland, and P.-Y. Barriat (2012), Constraining projections of summer Arctic sea ice, *Cryosphere*, **6**(6), 1383–1394, doi:10.5194/tc-6-1383-2012.
- Notz, D., and J. Stroeve (2016), Observed Arctic sea-ice loss directly follows anthropogenic CO<sub>2</sub> emission, *Science*, **354**(6313), 747–750, doi:10.1126/science.aag2345.
- Overland, J. E., and M. Wang (2013), When will the summer Arctic be nearly sea ice free?, *Geophys. Res. Lett.*, **40**, 2097–2101, doi:10.1002/grl.50316.
- Parker, W. S. (2006), Understanding pluralism in climate modeling, *Found. Sci.*, **11**(4), 349–368.
- Parker, W. S. (2009), Confirmation and adequacy-for-purpose in climate modelling, *Aristot. Soc. Suppl. Vol.*, **83**(1), 233–249, doi:10.1111/j.1467-8349.2009.00180.x.
- Pennell, C., and T. Reichler (2011), On the effective number of climate models, *J. Clim.*, **24**(9), 2358–2367, doi:10.1175/2010JCLI3814.1.
- Rausser, F., P. Gleckler, and J. Marotzke (2015), Rethinking the default construction of multimodel climate ensembles, *Bull. Am. Meteorol. Soc.*, **96**(6), 911–919, doi:10.1175/BAMS-D-13-00181.1.
- Reichler, T., and J. Kim (2008), How well do coupled models simulate today's climate?, *Bull. Am. Meteorol. Soc.*, **89**(3), 303, doi:10.1175/BAMS-89-3-303.
- Sanderson, B. M., and R. Knutti (2012), On the interpretation of constrained climate model ensembles, *Geophys. Res. Lett.*, **39**, L16708, doi:10.1029/2012GL052665.

- Sanderson, B. M., R. Knutti, and P. Caldwell (2015a), A representative democracy to reduce interdependency in a multimodel ensemble, *J. Clim.*, *28*(13), 5171–5194, doi:10.1175/JCLI-D-14-00362.1.
- Sanderson, B. M., R. Knutti, and P. Caldwell (2015b), Addressing interdependency in a multimodel ensemble by interpolation of model properties, *J. Clim.*, *28*(13), 5150–5170, doi:10.1175/JCLI-D-14-00361.1.
- Screen, J. A., and J. A. Francis (2016), Contribution of sea-ice loss to Arctic amplification is regulated by Pacific Ocean decadal variability, *Nat. Clim. Change*, *6*(9), 856–860, doi:10.1038/nclimate3011.
- Sherwood, S. C., S. Bony, and J.-L. Dufresne (2014), Spread in model climate sensitivity traced to atmospheric convective mixing, *Nature*, *505*(7481), 37–42, doi:10.1038/nature12829.
- Swart, N. C., J. C. Fyfe, E. Hawkins, J. E. Kay, and A. Jahn (2015), Influence of internal variability on Arctic sea-ice trends, *Nat. Clim. Change*, *5*(2), 86–89, doi:10.1038/nclimate2483.
- Tebaldi, C., and R. Knutti (2007), The use of the multi-model ensemble in probabilistic climate projections, *Philos. Trans. R. Soc. London, Ser. A*, *365*(1857), 2053–2075, doi:10.1098/rsta.2007.2076.
- Tebaldi, C., L. O. Mearns, D. Nychka, and R. L. Smith (2004), Regional probabilities of precipitation change: A Bayesian analysis of multimodel simulations, *Geophys. Res. Lett.*, *31*, L24213, doi:10.1029/2004GL021276.
- Watterson, I. G. (1996), Non-dimensional measures of climate model performance, *Int. J. Climatol.*, *16*(4), 379–391, doi:10.1002/(SICI)1097-0088(199604)16:4<379::AID-JOC18>3.0.CO;2-U.
- Waugh, D. W., and V. Eyring (2008), Quantitative performance metrics for stratospheric-resolving chemistry-climate models, *Atmos. Chem. Phys.*, *8*(18), 5699–5713, doi:10.5194/acp-8-5699-2008.
- Weigel, A. P., R. Knutti, M. A. Liniger, and C. Appenzeller (2010), Risks of model weighting in multimodel climate projections, *J. Clim.*, *23*(15), 4175–4191, doi:10.1175/2010JCLI3594.1.
- Wenzel, S., P. M. Cox, V. Eyring, and P. Friedlingstein (2014), Emergent constraints on climate-carbon cycle feedbacks in the CMIP5 Earth system models, *J. Geophys. Res. Biogeosci.*, *119*, 794–807, doi:10.1002/2013JG002591.
- Wenzel, S., V. Eyring, E. P. Gerber, and A. Y. Karpechko (2016a), Constraining future summer austral jet stream positions in the CMIP5 Ensemble by process-oriented multiple diagnostic regression, *J. Clim.*, *29*(2), 673–687, doi:10.1175/JCLI-D-15-0412.1.
- Wenzel, S., P. M. Cox, V. Eyring, and P. Friedlingstein (2016b), Projected land photosynthesis constrained by changes in the seasonal cycle of atmospheric CO<sub>2</sub>, *Nature*, *538*(7626), 499–501, doi:10.1038/nature19772.
- Zubler, E. M., A. M. Fischer, F. Fröb, and M. A. Liniger (2015), Climate change signals of CMIP5 general circulation models over the Alps—Impact of model selection, *Int. J. Climatol.*, doi:10.1002/joc.4538.